NILE UNIVERSITY

Market Basket Analysis using Apriori, FP-Growth, and Eclat

A CSCI467s Project Report

By

Name	ID
Eslam Ahmed Mohamed	202000039
Mariam Barakat	202000210
Reem Amgad Mostafa	202000438
Samaa Maged Ahmed	202000597
Abdelrahman Yasser Saeed	202000933

Submitted in partial fulfilment of the requirements.

for CSCI467s Project

To Dr. Walaa Medhat

28 May 2023

Market Basket Analysis using Apriori, FP-Growth, and Eclat

Eslam Ahmed, Mariam Barakat, Samaa Maged, Abdelrahman Yasser and Reem Amgad

Department of ITCS, Nile University

Sheikh Zayed, Egypt

Abstract **- Market Basket Analysis (MBA) is a widespread data mining approach for uncovering patterns and relationships between items that are often purchased together. The goal of this project is to apply MBA to a collection of relational data from a grocery store chain called *Instacart* to compare the usage of Apriori, FP-Growth and Eclat. Moreover, the paper analyses the dataset using PowerBI to extract information and insight. The findings of this study may be applied to consumer happiness by providing enough stock of frequently used items and help in the process of having worthwhile promotions.**

I. INTRODUCTION

Market Basket Analysis (MBA) has become a prevalent data mining approach for uncovering patterns and relationships among items frequently purchased together. By analyzing transactional data, MBA provides valuable insights into customer behavior and helps businesses optimize their strategies. The objective of this project is to apply MBA to a rich dataset obtained from Instacart, a prominent grocery store chain. We aim to compare the performance of three widely used MBA algorithms: Apriori, FP-Growth, and Eclat. These algorithms will be employed to extract frequent item sets and association rules from the dataset, revealing meaningful associations between products. Additionally, we will utilize PowerBI, a powerful data visualization tool, to extract further information and insights from the dataset.

By leveraging the findings of this study, businesses can gain a deeper understanding of consumer preferences and purchasing patterns. This knowledge can be utilized to improve inventory management by ensuring the availability of frequently purchased items, thereby enhancing customer satisfaction. Furthermore, the identification of strong associations between products can aid in the development of effective promotional strategies, facilitating targeted marketing campaigns and personalized recommendations. Overall, this project seeks to harness the power of MBA and data visualization techniques to drive actionable insights and improve decision-making processes in the grocery retail industry.

II. LITERATURE REVIEW

A. *Association Rules*

Association rules is a type of rule-based machine learning technique that discovers the patterns between different categories. To illustrate, association rules can be used to figure out hidden patterns between pairs of products purchased together [1]. This can be valuable information as shops can place items in ways that maximize sales.

There are important parameters that concern the association rules, namely, support, confidence, and lift [2]. To explain, support is the frequency of an item divided by the number of transactions as shown in eq.1. The X and Y are variables representing different items, and the N is the total number of transactions. To add, confidence as shown in eq.2 calculates how often Y is purchased when item X is purchased [2]. Moreover, lift is as indicated in eq.3 it showcases the strength of association between two items [2]. Furthermore, the Leverage eq.4 measures the difference between the observed frequency of two items occurring together and the frequency that would be expected if the items were independent [2]. Lastly, the conviction eq.5 provides a quantifiable measure of how much the presence or absence of the antecedent influences the consequent. Higher conviction values indicate a stronger dependency between the antecedent and consequent, indicating a more significant association.

$$Support(X \to Y) = \frac{Freq(X,Y)}{N} = P(X \cup Y) \quad (1)$$

$$Confidence(X \to Y) = \frac{Freq(X,Y)}{Freq(x)} = P(Y \mid X) \quad (2)$$

$$Lift\ (X \rightarrow Y) = \frac{Confid(X \rightarrow Y)}{Freq(x)} = \frac{Confid(X,Y)}{Freq(x)} \quad (3)$$

$$Leverage(X \rightarrow Y) = Supp(X \rightarrow Y) - Supp(X)Supp(Y) \quad (4)$$

$$Conviction = \frac{1 - support(Y)}{1 - Confidence(X \rightarrow Y)} \quad (5)$$

To effectively run the association rules, two main steps must be implemented. Firstly, to filter out all frequent item sets that pass the minimum support count. Secondly, is to create association rules using rule-based ML algorithms like Apriori or FP-Growth [1].

B. *Apriori*

Apriori is an important algorithm in data mining. It is built on the fact that large datasets can be reduced by pruning infrequently purchased items. Therefore, the output would have the essence of associated items [1].

Apriori Algorithm steps [2]:

1. Choose the minimum support level and search for items that meet this support level.
2. Utilize the frequent item set and generate candidate itemset of length k+1 by joining frequent item sets of length k.
3. Filter candidate item set using the support level.
4. Repeat steps 2 and 3 until no new frequent item sets appear.
5. Generate association rules using the minimum confidence level.

C. *FP – Growth*

Another data mining algorithm is FP-Growth. This algorithm has mainly two steps. Firstly, an FP-tree is constructed. This is a compressed representation of the data but in a tree format. Secondly, the algorithm recursively mines the FP-tree to generate all frequent item sets.

FP-Growth Algorithm steps [2]:

1. Choose the minimum support level and search for items that meet this support level.
2. Sort the remaining frequent items in decreasing order of frequency.
3. Build the FP-tree by grouping transactions that share the same items as shown in fig.1.
4. Recursively mine the FP-tree by generating conditional patterns for each frequent item in the tree.
5. Combine the results of the previous step to obtain the complete set of frequent item sets.
6. Generate association rules using the minimum confidence level.

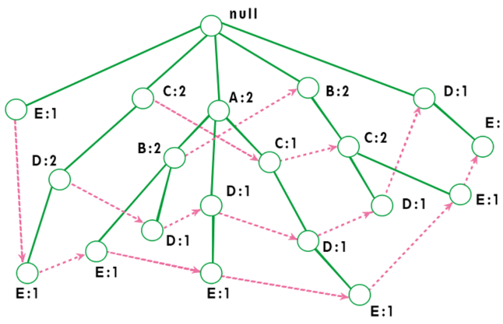


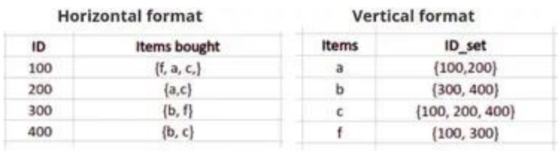
Fig. 1 FP-tree generated from itemset database [2].

D. *Eclat*

Eclat stands for Equivalence Class Clustering and Bottom-Up Lattice Traversal. This algorithm is built on the idea of Apriori, but it can be faster. The main difference is in representing the transactional data in a vertical format [3]. It should be mentioned that the Eclat also does not output metrics as Apriori and FP-Growth. In other words, the built-in Eclat does not give confidence and lift metrics.

Eclat Algorithms steps [3]:

1. List the Transaction ID set of each product by converting the transaction table or list of products bought by each user into a vertical format as shown in fig.2.
2. Filter with minimum support.
3. Compute the set of product pairs with the common transaction ID.
4. Filter out pairs that do not reach the minimum support.
5. Repeat until no new pairs are forming.



Horizontal format

ID	Items bought
100	{f, a, c,}
200	{a,c}
300	{b, f}
400	{b, c}

Vertical format

Items	ID_set
a	{100,200}
b	{300, 400}
c	{100, 200, 400}
f	{100, 300}

Fig. 2 Horizontal vs Vertical transactional table [4]

E. *Comparing Apriori, FP-Growth and Eclat*

The Apriori algorithm is simple to implement and has low memory utilization and execution time, it has some disadvantages compared to FP-Growth, particularly with larger datasets. However, FP-Growth is not good at generating strong candidates. [5] agree with these results and state that Apriori takes longer to implement and requires more memory than FP-Growth. FP-Growth is quicker than Apriori due to its divide and conquer strategy [6]. The Apriori algorithm scans the dataset multiple times to generate candidates, while FP-Growth scans it twice [7]. To minimize the disadvantages of these algorithms, studies suggest combining the outcomes of both Apriori and FP-Growth algorithms [1]. Furthermore, Eclat like Apriori is simple to implement, however it may work poorly with sparse data [4]. Moreover, Eclat scans the data only once making it sometimes faster than Apriori and FP-Growth [4]. Table 1 states the literature performance and memory complexity, the n is the number of transactions and m is Average number of items per transaction. The best algorithm to be used depends on the number of transactions and the sparsity of the data.

TABLE I
PERFORMANCE AND MEMORY COMPLEXITY COMPARISON

Algorithm	Average Case Complexity	Worst Case Complexity	Space Complexity
Apriori	$O(n^2 m)$	$O(2^m)$	$O(n^2 m)$
FP-Growth	$O(n\, m\, log(m))$	$O(2^m)$	$O(n\, m)$
Eclat	$O(n\, m\, (2^m))$	$O(2^m)$	$O(n\, m)$

III. METHODOLOGY

The methodology for this study focuses on utilizing Market Basket Analysis to predict customers' next purchased items. Specifically, two popular data mining algorithms, Apriori and FP-Growth, are employed for this purpose. The study utilizes the Instacart dataset as the primary data source.

To begin, the methodology involves preprocessing the Instacart dataset to ensure data quality and consistency. This includes handling missing values, data normalization, and transforming the dataset into a suitable format for market basket analysis.

Subsequently, it utilizes Market Basket Analysis, specifically employing the Apriori and FP-Growth algorithms, to predict customers' next purchased items using the Instacart dataset. By leveraging these data mining techniques, the study aims to provide valuable insights into customer behavior and enhance personalized marketing strategies.

A. Feature Description

Understanding the features thoroughly is the first step to being able to have an intuition of how to use the dataset. Firstly, the dataset consists of 6 tables, namely: Aisles, Departments, Products, Orders, Orders_Products_Prior, and Orders_products_Train. Below are tables describing features thoroughly.

TABLE III
AISLE TABLE

aisle_id	Unique id for each aisle
aisle	Aisle contents ex: prepared soups salads, specialty cheese, etc.

Table I describes the Aisle content, it connects each aisle with what the aisle contains.

TABLE IIIII
DEPARTMENT TABLE

department_id	Unique id for each department
department	Department contents ex: frozen, bakery, etc.

Table II describes the department table where the department_id corresponds to the department name.

TABLE IVV
PRODUCT TABLE

product_id	Unique id for each product
product_name	Product name for the specific id
aisle_id	The aisle id in which the product exists
department_id	The department id in which the product exists

Table III describes the products. Each product has its own product_id, which showcases the product name, the aisle_id describing the ID of the aisle it is present in. And finally, the department_id it is considered in.

TABLE V
ORDERS TABLE

order_id	Unique id for each order
user_id	The id of the user of the specific order
eval_set	Tells which set (prior, train, test) an order belongs
order_number	The number of the order
order_dow	the day of week the order was made
order_hour_of_day	the hour of the day the order was made
days_since_prior_order	The days since the previous order was made if this order is reordered

Table IV describes the orders table; it includes the order_id and the user_id who purchased this order. Also, eval_set differentiates whether this row is considered as a test or train data. Moreover, the order_number describes the number of products bought in the specific order_id. The order_dow, order_hour_of_day, and days_since_prior_order are the day of the week the order was made, the hour of purchase and days since the previous order was made respectively.

TABLE VI
ORDERS PRODUCTS PRIOR

order_id	The id of the previous order
product_id	The product id in the previous order
add_to_cart_order	Previous product was purchased in what order
reordered	If the customer has a previous order that contains the product (1 or 0)

For table IV, the order_id for testing is the id of the previous order of the same user. This is linked with a group of products_ids that were ordered. Lastly, reorder has a Boolean datatype that explains whether the product was reordered or not.

TABLE VII
ORDERS PRODUCTS TRAIN

order_id	The id of the previous order
product_id	The product id in the previous order
add_to_cart_order	Previous product was purchased in what order
reordered	If the customer has a previous order that contains the product (1 or 0)

For table VI, the order_id for training is the id of the previous order of the same user. This is linked with a group of products_ids that were ordered. Lastly, reorder has a Boolean datatype that explains whether the product was reordered or not.

B. Data Warehouse Schema

In this study, we employed a data warehousing approach to analyze the Instacart dataset. Data warehousing is a technique that involves collecting, organizing, and storing large volumes of data from various sources to support effective data analysis and decision-making processes. The primary goal of data warehousing is to provide a unified and structured view of data, enabling easy access, manipulation, and exploration.

To implement the data warehousing solution, we utilized PostgreSQL. PostgreSQL offers scalable storage and processing capabilities, making it suitable for handling large datasets. We designed the data warehouse using a snowflake schema, a popular multidimensional modeling technique.

The snowflake schema organizes data into a central fact table surrounded by multiple dimension tables, forming a hierarchical structure. This schema allows for efficient and flexible querying by separating data into granular dimensions and reducing data redundancy. By adopting the snowflake schema in our data warehouse, we were able to organize the Instacart dataset into a logical and optimized structure for analysis. Figure 3 illustrates this.

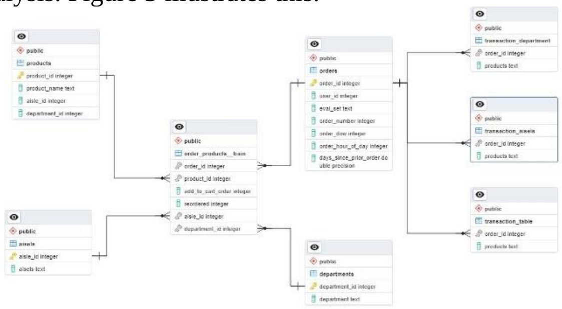


Fig. 3 Star Schema for Instacart data

C. Market Basket Analysis

The methodology for conducting market basket analysis consisted of the following steps:

1. Data Preprocessing: We started by preparing the data for analysis. We created a file that contained transaction IDs with corresponding lists of items purchased in each transaction. This data format is commonly used for market basket analysis.

2. Algorithm Implementation: We implemented three popular association rule mining algorithms, Apriori, FP-growth and Eclat, to extract frequent item sets and generate association rules. For the Apriori and FP-growth algorithms, we utilized the mlxetend library, which provides state-of-the-art implementations of these algorithms in Python. Additionally, we implemented the Eclat algorithm using the pyECLAT library. It must be noted that a support of 0.01 is used as it was the best option that yields to worthwhile association rules. This was chosen upon trial and error.

3. Association Rule Visualization: After running Apriori and FP-growth, we generated 3D and 2D

graphs to visualize the resulting association rules. These graphs provided a visual representation of the relationships and associations between different items in the dataset. This allowed for easier comparison and interpretation of the discovered patterns.

4. Time Measurement: To evaluate the performance of the algorithms, we measured and recorded the execution time for each algorithm. This step helped us understand the efficiency and computational requirements of Apriori, FP-growth, and Eclat in our specific dataset.

By following this methodology, we were able to extract association rules using Apriori, FP-growth, and Eclat algorithms, visualize the results, and measure the execution time. These steps allowed us to gain insights into the relationships and patterns in the market basket data and compare the performance of the different algorithms in terms of runtime and rule generation.

D. Algorithm Evaluation

In this study, we evaluated the performance of two popular association rule mining algorithms, FP-Growth and Apriori, and compared them based on key measures such as support, confidence, leverage, lift, and conviction.

To conduct the evaluation, we utilized a dataset containing transactional data from a grocery store chain. This dataset served as the input for the association rule mining algorithms. We applied FP-Growth and Apriori algorithms to discover frequent item sets and generate association rules based on predefined thresholds.

For each algorithm, we calculated the support, confidence, leverage, lift, and conviction measures for the generated association rules. The support measure represents the proportion of transactions that contain both the antecedent and consequent items. Confidence measures the conditional probability of finding the consequent given the antecedent. Leverage quantifies the difference between the observed and expected co-occurrence of the antecedent and consequent. Lift assesses the strength of the association by comparing the observed support with the expected support under independence. Conviction provides a measure of how much the absence of the antecedent influences the absence of the consequent.

Through this evaluation, we were able to compare the performance of FP-Growth and Apriori algorithms in terms of rule quality, significance, and effectiveness in capturing interesting associations within the dataset. However, it's important to note that Eclat algorithm, which is known for its efficiency in terms of memory usage, focuses primarily on support metrics and was not directly comparable in other aspects such as confidence, leverage, lift, and conviction.

By analyzing and comparing these measures across FP-Growth and Apriori, we gained valuable insights into the strengths and limitations of each algorithm, helping us make informed decisions about their applicability in our association rule mining tasks within the grocery store dataset.

IV. RESULTS

In the results section, we present the insightful findings obtained from the market basket analysis and the visualization of data using PowerBI. Through the market basket analysis, we discovered meaningful patterns and associations among items frequently purchased together by customers. These findings shed light on customer preferences and shopping behaviors, enabling businesses to make informed decisions. For instance, we identified strong associations between certain products, indicating that customers who bought one item were highly likely to purchase another. Such knowledge can be leveraged to optimize inventory management, design effective product placements, and develop targeted marketing strategies to enhance customer satisfaction and drive sales.

Additionally, PowerBI played a crucial role in uncovering deeper insights from the market basket data. By utilizing its visualization and data exploration capabilities, we created interactive dashboards and reports that revealed valuable trends and patterns. We gained a comprehensive overview of customer behavior, including top-selling items, customer segmentation based on purchasing habits, and seasonal variations in product demand. These insights provide a holistic understanding of the data and empower businesses to optimize their strategies, improve customer satisfaction, and drive growth by tailoring their offerings and marketing campaigns to align with customers' preferences and buying patterns.

A. Market Basket Analysis insights

To start, a code was implemented to figure out the top purchased 25 products to be able to have a base of frequencies when implementing the association rules, figure 3 illustrates that. It can be inferred that this shop has mostly healthy choices, or it is known for its healthy variations.

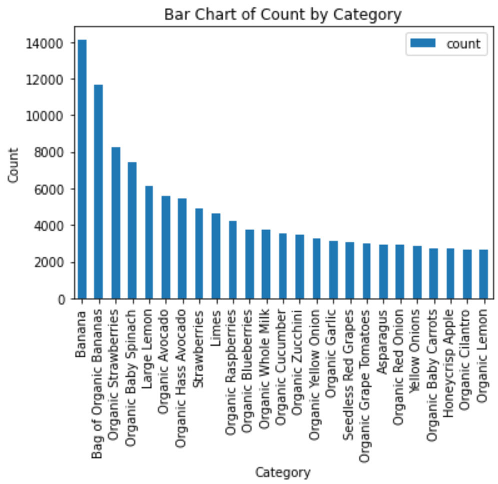


Fig. 4 Top 25 products

Figure 4 is a representation of confidence as it increases, this is to show that the number of associations decrease steeply as the confidence threshold needed increases.

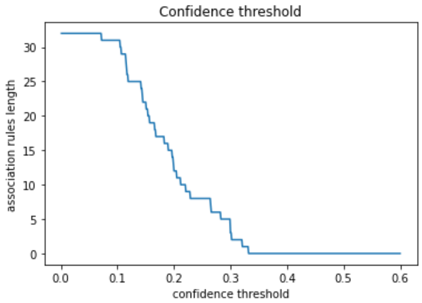


Figure 5 Confidence Thresholds

To be able to compare without having a lot of association rules. In figure 5 it is shown how the confidence was chosen from the figure. As the 30 products were desired, the confidence threshold chosen was 0.1.

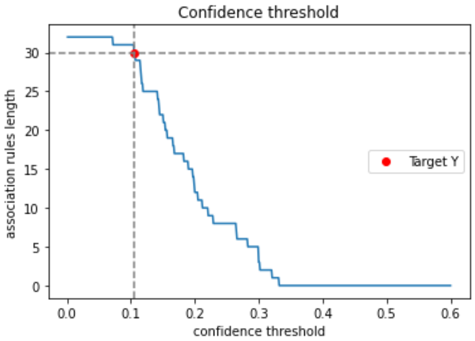


Fig. 6 Choosing 30 products using Confidence threshold.

Figure 6 compares the lift, confidence, and support of products visually in Apriori. As the support increased the more repeated those were together. The higher the confidence the stronger the association between the antecedent and consequent in an association rule. Lastly, the higher the lift measures the strength of association between the antecedent and consequent independent of their individual support. Figure 7 shows the 3D equivalent.

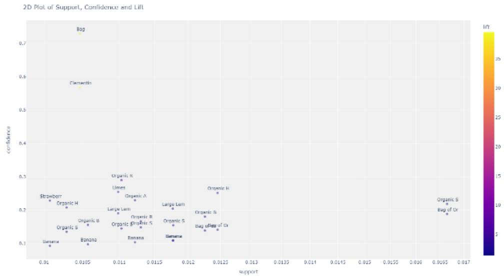


Fig. 7 2D Plot of Support, Confidence, and lift for Apriori

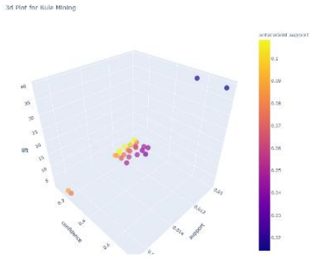


Fig. 8 3D Plot of Support, Confidence, and lift for Apriori

Figure 8 compares the lift, confidence, and support of products visually in FP-growth. As the support increased the more repeated those were together. The higher the confidence the stronger the association between the antecedent and consequent in an association rule. Lastly, the higher the lift measures the strength of association between the antecedent and consequent independent of their individual support. Figure 9 shows the 3D equivalent.

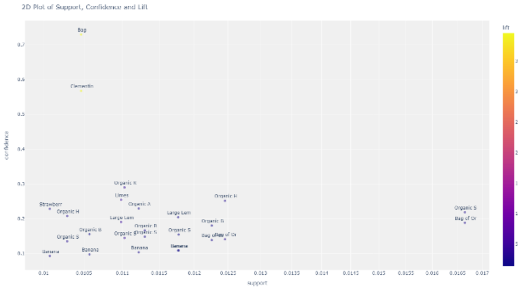


Fig. 9 2D Plot of Support, Confidence, and lift for FP-Growth

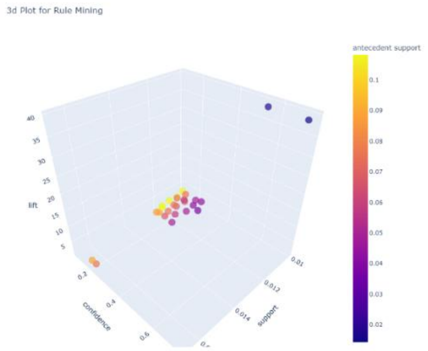


Fig. 10 3D Plot of Support, Confidence, and lift for FP-Growth

It can be noted that the shape of the 2D and 3D plots seems very similar however they are different products.

For the performance complexity, figure 10 shows how many seconds each algorithm took to run. Notice that Eclat took too long, therefore its transaction was decreased until it was a reasonable time, which is 10 transactions for Eclat to run on. This result was unexpected, especially for Eclat. This may be because mlxtend used in Apriori and FP-growth has used advanced generators that made processing faster. FP-Growth is considered the best for this dataset.

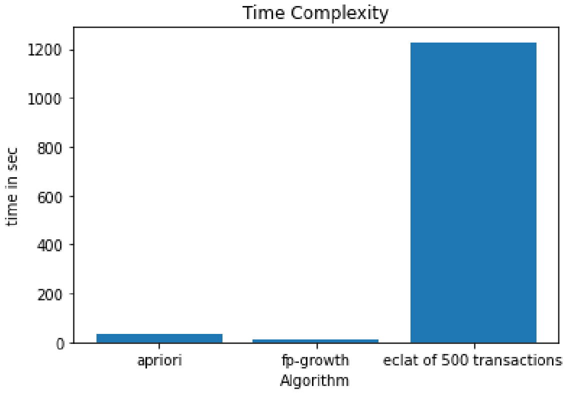


Fig. 11 Time complexity of the three algorithms

In this study, we utilized the Apriori algorithm and association rules to explore potential collaborations between departments and optimize product placement within aisles. The objective was to identify departments that should collaborate more closely based on their co-occurrence in customer transactions, and to determine which products should be placed near each other within aisles.

Using the Apriori algorithm, we extracted frequent item sets representing department combinations that frequently appeared together in customer transactions seen in fig. [12 – 13]. By applying association rules to these item sets, we generated insights on departments that showed strong associations, indicating potential opportunities for collaboration. This information can guide decision-making in terms of resource allocation, cross-promotions, and overall store layout optimization.

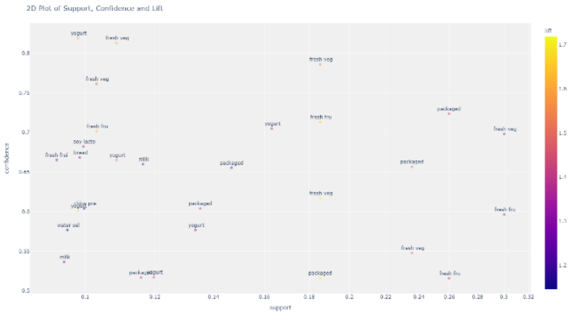


Fig. 12 2D Association rules of departments

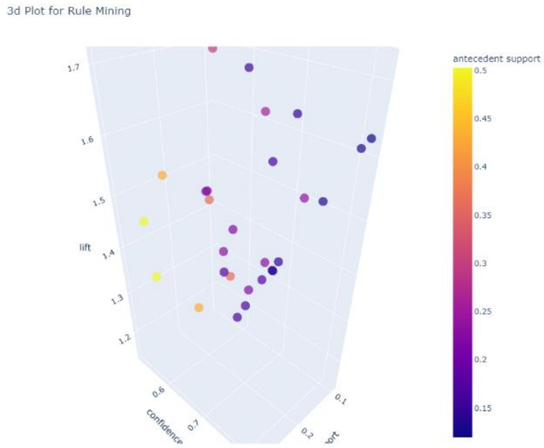


Fig. 13 3D Association rules of departments

Additionally, we analyzed association rules specific to product placements within aisles. By examining the co-occurrence of products in customer transactions, we identified associations between products that are often purchased together. These associations can inform merchandising strategies, suggesting which products should be positioned near each other within aisles to enhance customer convenience and increase sales. The results are illustrated in fig. [14 - 15].

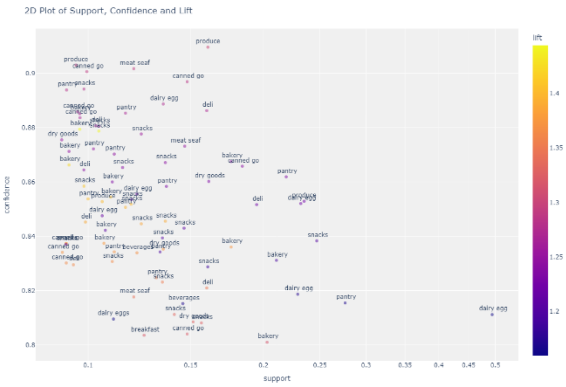


Fig. 14 2D Association rules of Aisles

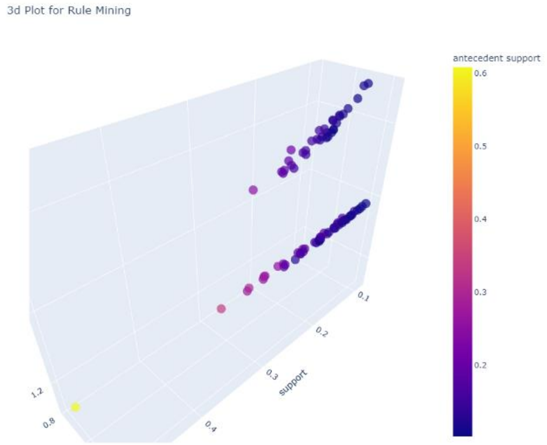


Fig. 15 3D Association rules of Aisles

V. CONCLUSION

In conclusion, our analysis of market basket data using different association rule mining algorithms has provided valuable insights into item relationships and purchasing patterns. Among the algorithms we evaluated, FP-growth emerged as the most suitable approach for this dataset.

FP-growth demonstrated superior performance in terms of both computational efficiency and memory usage compared to Apriori and Eclat. It effectively handled the large volume of transaction data and generated frequent item sets and association rules with high support and confidence. The compact representation of the FP-tree data structure allowed for efficient mining of item sets, making it well-suited for large-scale market basket analysis.

On the other hand, Eclat, despite being a popular algorithm, did not perform as expected in our analysis. The sparse nature of the data may have hindered its ability to extract meaningful associations between items. The lack of support for additional metrics such as lift, and conviction limited our ability to compare Eclat with the other algorithms comprehensively.

Overall, FP-growth proved to be a powerful and efficient algorithm for market basket analysis, enabling us to uncover valuable insights into item associations and purchasing patterns. Further research and experimentation could explore ways to improve Eclat's performance on sparse datasets and enhance its suitability for market basket analysis.

REFERENCES

[1] D. Alcan, K. Ozdemir, B. Ozkan, A. Y. Mucan, and T. Ozcan, “A comparative analysis of apriori and FP-growth algorithms for market basket analysis using multi-level association rule mining,” in *Lecture Notes in Management and Industrial Engineering*, Cham: Springer Nature Switzerland, 2023, pp. 128–137.

[2] S. Halim, T. Octavia, and C. Alianto, “Designing facility layout of an amusement arcade using market basket analysis,” *Procedia Comput. Sci.*, vol. 161, pp. 623–629, 2019.

[3] J. Korstanje, “The eclat algorithm,” *Towards Data Science*, 29-Sep-2021. [Online]. Available: https://towardsdatascience.com/the-eclat-algorithm-8ae3276d2d17. [Accessed: 27-May-2023].

[4] M. S.Mythili and A. R. Mohamed Shanavas, “Performance evaluation of apriori and FP-growth algorithms,” *Int. J. Comput. Appl.*, vol. 79, no. 10, pp. 34–37, 2013.

[5] A. K. Singh, A. Kumar, and A. K. Maurya, “An empirical analysis and comparison of apriori and FP- growth algorithm for frequent pattern mining,” in *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, 2014.

[6] R. Garg and P. Gulia, “Comparative study of frequent itemset mining algorithms apriori and FP growth,” *Int. J. Comput. Appl.*, vol. 126, no. 4, pp. 8–12, 2015.